

MAT 2379, Introduction to Biostatistics
Section 9.1. Part 3: Transformation of variables

Sometimes, it is useful to transform a data set into a new data set, using the same rule for transforming each data point.

A **linear transformation** of a variable X is a new variable of the form:

$$X' = aX + b$$

Example 1. The following data set gives the temperatures for 4 women in degrees Celsius:

$$x_1 = 36.23, \quad x_2 = 36.41 \quad x_3 = 36.77 \quad x_4 = 36.15$$

The sample mean, sample variance, and sample standard deviation are:

$$\bar{x} = 36.39, \quad s^2 = 0.076, \quad s = 0.2757$$

We would like to express the data in degrees Fahrenheit. For this, we use the linear transformation:

$$X' = 1.8X + 32$$

The new data points are:

$$x'_1 = 97.214, \quad x'_2 = 97.538 \quad x'_3 = 98.186 \quad x'_4 = 97.07$$

The sample mean, sample variance, and sample standard deviation of the new sample are:

$$\bar{x}' = 97.502, \quad (s')^2 = 0.246, \quad s' = 0.496$$

Note that:

$$\bar{x}' = 1.8\bar{x} + 32, \quad (s')^2 = (1.8)^2 s^2, \quad s' = 1.8s$$

In general, if x_1, \dots, x_n are observations from a variable X , and x'_1, \dots, x'_n are the linearly transformed data points:

$$x'_i = ax_i + b, \quad i = 1, \dots, n$$

then

$$\bar{x}' = a\bar{x} + b, \quad (s')^2 = a^2 s^2, \quad s' = |a|s$$

Note: In the case of non-linear transformations (e.g. $X' = \sqrt{X}$, $X' = e^X$, etc.), the previous property does not hold.

A **logarithmic transformation** of a variable X is a new variable of the form:

$$Y = \ln(X)$$

Consider the logarithmic measurements $y_i = \ln(x_i), i = 1, \dots, n$. The *geometric mean* of the sample x_1, \dots, x_n is defined by:

$$g = e^{\bar{y}}, \quad \text{where} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

The *geometric standard deviation* of the sample x_1, \dots, x_n is defined by:

$$e^{s_y}, \quad \text{where} \quad s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}.$$